PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(b)(2)

Docket No. 02163-0120P

INVENTOR(s)

LAST NAME	FIRST NAME	M.I.	RESIDENCE (city & either state or foreign country)
BARNHILL	STEPHEN	D.	19 Mad Turkey Crossing Savannah, Georgia 31411

TITLE OF THE INVENTION (280 characters max)

A METHOD FOR DISCOVERING KNOWLEDGE USING SUPPORT VECTOR MACHINES

CORRESPONDENCE ADDRESS

JONES & ASKEW, LLP
37th Floor
191 Peachtree Street
Atlanta, Georgia 30303-1769

Attn: James Dean Johnson

ENCLOSED APPLICATION PARTS *(check all that apply)*

☒ Specification *Number of Pages* 6

☐ Drawing(s) *Number of Sheets*

☐ Provisional Application Filing Fee

☐ Small Entity Statement

☐ Other (specify)

METHOD OF PAYMENT

☐ A check is enclosed to cover the Provisional Application filing fee.

FILING FEE: $

The invention was not made by an agency of the U.S. Government nor under a contract with an agency of the U.S. Government.

Respectfully submitted,

SIGNATURE: Mary Anthony Merchant Date: May 1, 1998

TYPED OR PRINTED NAME: Mary Anthony Merchant, Ph.D. Reg. No. 39,771

☐ Additional inventors are being named on separately numbered sheets attached hereto

"Express Mail" Mailing Label number EM515355193US Date of Deposit: May 1, 1998

PENGAD-Bayonne, N.J.
EXHIBIT
1

A Method for Discovering Knowledge Using Support Vector Machines.

Inventor - Stephen D. Barnhill, M.D.

Technical Field:

The present invention relates to methods for extracting desired data from databases. More particularly, the present invention relates to a method for extracting desired data from databases from generated and collected sets of data, large or small, relating to humans, animals, viruses, and bacteria, as well as accounting data, stock and commodity market data, and insurance data, in order to effectively classify subgroups by virtue of a computational index *(CompuDex™)*. The present invention creates an effective method for multi-dimensional function estimation and resulting ***CompuDex™*** that can be applied to a wide range of problems including pattern recognition, function approximation, regression estimation, molecular patterning, proportionality estimations and signal processing.

The present invention further relates to a computer assisted method for classifying subgroups utilizing pre-processed ***IntelliData™*** (Intelligent Data created by pre-processing techniques utilized specifically as part of the present invention). The ***IntelliData™*** is then entered into one or more Support Vector Machines which generates an optimal hyperplane algorithm. This optimal hyperplane algorithm is then converted by one or more post-processing steps into a ***CompuDex™*** (a single valued computationally derived numerical classifier) for interpretation by a human. In summary, the present invention begins with raw data and using support vector machines then concludes with a single valued computationally derived numerical classifier ready for human interpretation.

In the preferred embodiment of the present invention, the method is used to classify individual subgroups, based on pattern recognition techniques, from any combination of raw data. Examples of the usefulness of this procedure could be demonstrated in 1.) genetics in general and the genome project specifically, 2.) diagnostics, 3.) evaluation of managed care efficiency, 4.) therapeutic decisions and follow up, 5.) appropriate therapeutic triage, 6.) pharmaceutical development techniques, 7.) discovery of molecular structures, 8.) prognostic evaluations, 9.) medical informatics, 10.) billing fraud, 11.) inventory control, and 12.) stock evaluations and predictions, 13.) commodity evaluations and predictions, and 14.) insurance probability estimates.

In another preferred embodiment, the invention includes a system to receive data from a remote data transmitting station for processing through the invention and transmit the results to the same or some other remote data receiving station.

Background of the Invention:

Knowledge discovery is the most desirable end-product of data collection. The last decade has brought forward an explosive growth in our capabilities to both generate, collect, and store vast amounts of data. While database technology has provided the basic tools for the efficient storage and collection of large data sets, the issue of how to help humans understand and analyze large bodies of data remains a difficult and unsolved problem.[1] In order to deal with this data glut, a new generation of intelligent tools for automated knowledge is needed.[2]

Exhibit 1

For example, there are huge scientific databases such as in the Human Genome Project which include gigabytes of data on the human genetic code and much more is expected.[3] Such volumes of data clearly overwhelm the traditional manual methods of data analysis, such as spread sheets and ad-hoc queries. Those methods can create informative reports from data, but do not have the capacity to discover the knowledge contained in the data. A significant need exists for a new generation of techniques and tools with the ability to intelligently and automatically assist humans in analyzing the mountains of data and finding patterns of useful knowledge.[4]

Likewise, using traditionally accepted reference ranges and standards for interpretation, it is often impossible for humans to identify patterns of useful knowledge even with very small amounts of data.

This invention in part utilizes Support Vector Machines. The Support Vector Machine implements the following idea: it maps the input vectors into high dimensional feature space through non-linear mapping, chosen *a priori*. In this high dimensional feature space, an optimal separating hyperplane is constructed. This Optimal Hyperplane Classifier Algorithm separates the various classes of interests.

The dimensionally of the feature space will be will be huge. For example, to construct a polynomial of degree 4 or 5 in a 200 dimensional space it is necessary to construct hyperplanes in a billion dimensional feature space. This curse of dimensionality can be solved by constructing the Optimal hyperplane. If it happens that in the high dimensional input space one can construct a separating hyperplane with a small value, the VC dimension of the corresponding element of the structure will be small, and therefore the generalization ability of the constructed hyperplane will be high.

If the training vectors are separated by the Optimal hyperplane (or generalized Optimal hyperplane), then the expectation value of the probability of committing an error on a test example is bounded by the examples in the training set.

This bound depends neither on the dimensionality of the space, nor on the norm of the vector of coeefficients, nor on the bound of the norm of the input vectors. Therefore, if the Optimal hyperplane can be constructed from a small number of support vectors relative to the training set size, the generalization ability will be high—even in infinite-dimensional space.

The problems with Back-Propagation Neural Network approach such as:

1.) The empirical risk fuctional has many local minima. Standard optimization proceedures guarantee convergence to one of them. The quality of the obtained solution depends on many factors, in particular on the initialization of the weight matrices.
2.) The convergence of the gradient based method is rather slow.
3.) The sigmoid function has a scaling factor which affects the quality of the approximation.

prevent neural networks from being well controlled learning machines.

These shortcomings of neural networks are overcome using Support Vector Machines by constructing the Optimal hyperplane. Support Vector Machines are described in detail in ***The Nature of Statistical Learning Theory*** by Vladimir Vapnik and are incorporated herein by reference in their entirety.

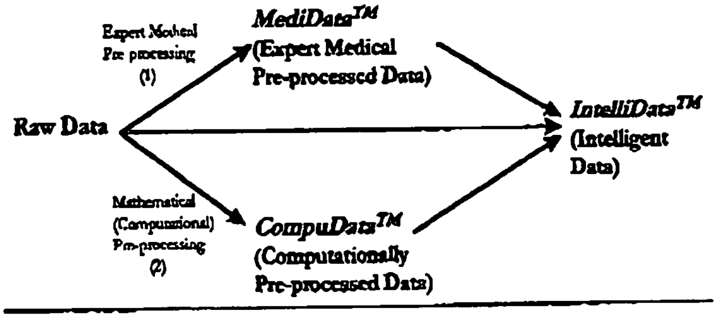
Summary of Invention:

The present invention is an apparatus and a process for classifying subgroups, based on pattern recognition techniques, utilizing ***IntelliData™*** (Intelligent Data created by pre-processing techniques utilized specifically as part of the present invention). The ***IntelliData™*** is then entered into one or more Support Vector Machines which generates an optimal hyperplane algorithm. This optimal hyperplane algorithm is then converted by one or more post-processing steps into a ***CompuDex™*** (a single valued computationally derived numerical classifier) for interpretation by a human.

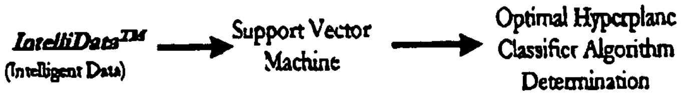
Generally, this objective is accomplished by performing the following steps:

1. Collect data in it's original and/or natural form.
2. Optionally apply expert medical pre-processing techniques to derive ***MediData™*** (Data derived from applying expert medical information to raw data).
3. Optionally apply mathematical (computational) pre-processing techniques to derive ***CompuData™***. (Data derived from applying mathematical (computational) informat on to raw data).
4. Combine the results of ***MediData™*** and ***CompuData™*** with the original raw data to create ***IntelliData™***.
5. Utilize the created ***IntelliData™*** as input into one or more support vector machines.
6. Generate an optimal hyperplane classifier algorithm.
7. Apply mathematical post-processing techniques to create a ***CompuDex™*** result.

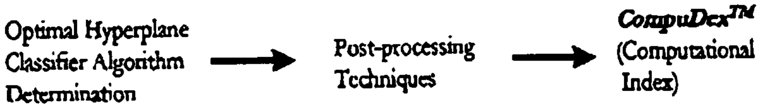
Step 1 - Creation of Intelligent Data - *IntelliData™*


Expert Medical Pre processing (1)
MediData™ (Expert Medical Pre-processed Data)
Raw Data
IntelliData™ (Intelligent Data)
Mathematical (Computational) Pre-processing (2)
CompuData™ (Computationally Pre-processed Data)

Step 2 - Determination of the Optimal Hyperplane Classifier Algorithm



Step 3 - Creation of a Mathematical CompuDex™ (Computational Index) for Human Interpretation



More detail performing the steps in the invention is as follows:

1. **Collect data in it's original and/or natural form.**

   This initial step involves generating and collecting any given set of data that may contain information which is not immediately apparent and needs to be evaluated to identify any patterns of useful knowledge.

2. **Optionally apply expert medical pre-processing techniques to derive *MediData™*.**
   This step actually creates an additional new set of input data (input vectors).

   This next step in the invention involves the option of application of expert medical pre-processing techniques to the raw data to create an additional set of input data known as *MediData™*. Examples of expert medical pre-processing steps include but are not limited to the following:
   A. Association with known standard reference ranges.
   B. Physiologic Truncation
   C. Physiologic Combinations
   D. Biochemical Combinations
   E. Application of Heuristic Rules
   F. Diagnostic Criteria Determinations
   G. Clinical Weighting Systems
   H. Diagnostic Transformations
   I. Clinical Transformations
   J. Application of Expert Knowledge
   K. Labeling Techniques
   L. Application of other Domain Knowledge
   M. Bayesian Network Knowledge

3. **Optionally apply mathematical (computational) pre-processing techniques to derive *CompuData™*.**
   This step actually creates an additional new set of input data (input vectors).

   This next step in the invention involves the option of application of mathematical (computational) pre-processing techniques to the raw data to create an additional set of input data known as *CompuData™*. Examples of mathematical (computational) pre-processing steps include but are not limited to the following:

A. Labeling
B. Binary Conversion
C. Logarithmic Transformation
D. Sine Transformation
E. Cosine Transformation
F. Tangent Transformation
G. Cotangent Transformation
H. Clustering
I. Summarization
J. Scaling
K. Probabilistic Analysis
L. Significance Testing
M. Strength Testing
N. Search for 2-D Regularities
O. Identify Equivalence Relations
P. Apply Contingency Tables
Q. Apply Graph Theory Principles
R. Create Vectorizing Maps
S. Multiplication
T. Division
U. Addition
V. Subtraction
W. Application of Polynomial Equations
X. Application of Basic and Complex Statistics
Y. Identify Proportionality's
Z. Discriminatory Power Determination
AA. Apply Combinations of the Above Simultaneously

4. Combine the results of *MediData™* and *CompuData™* with the original raw data to create *IntelliData™*.
This step actually creates an additional new set of input data (input vectors).

This step of the invention combines the attributes (vectors) of the raw data, the *MediData™* and the *CompuData™* to create an additional new set of input data (input vectors) called *IntelliData™* to be fed into the Support Vector Machines for high dimensional computation and mapping.

5. Utilize the created *IntelliData™* as input into one or more support vector machines.

This step of the invention utilizes the original raw data (vectors) along with the option of using newly created vectors *MediData™*, *CompuData™*, and *IntelliData™* to assist in providing smarter data to the Support Vector Machine to allow for better computation and high dimensional mapping in the creation of the optimal hyperplane algorithm

6. Generate an optimal hyperplane classifier algorithm

This step of the invention uses one or more Support Vector Machines to determine the optimal hyperplane classifier algorithm. The kernal of the Support Vector Machine can be a polynomial kernal, a radial bias classifier kernal, a neural network kernal, or any other type of kernal that satisfies the Mercer Condition.

To construct the Optimal Hyperplane, one has to separate the vectors of the training set belonging to two different classes using the hyperplane with the smallest norm of coefficients. The Support Vector Machine implements the following idea: it maps the input vectors into high dimensional feature space through some non-linear mapping, chosen *a priori*. In this space, an optimal separating hyperplane is constructed. This Optimal Hyperplane Classifier Algorithm separates the various classes of interest.

7. **Apply mathematical post-processing techniques to create a *CompuDex™* result.**

This step of the invention takes the Optimal Hyperplane Classifier Algorithm and optionally apply post-processing techniques to create a ***CompuDex™*** (a computational index) which can then be interpreted by a human. Examples of Post-processing steps include but are not limited to the following:

A. Reference Range Determinations
B. Scaling Techniques (linear and non-linear)
C. Transformations (linear and non-linear)
D. Probability Estimations

Conclusion:

This invention is a successful method for extracting and manipulating data in data sets, large or small, using data pre-processing steps to create IntelliData™ (intelligent data), which is then analyzed in high dimensional space using one or more support vector machines; the results of which are then subjected to post-processing steps to create a single valued numerical classifier, CompuDex™ (a computational index), which can then be easily interpreted by a human.